Three Big Trends

Explosion of Data

- 90% Unstructured
- Video & Image Content
- Needs Higher Throughput & Real-Time Computing

© Copyright 2018 Xilinx
Three Big Trends

Dawn of AI

> Adoption Across All Industries
> Injecting New Intelligence into Apps
> From Endpoints to Edge to Cloud
Three Big Trends

Computing After Moore’s Law

- Heterogeneous Computing with Accelerators
- Breadth of Apps and AI Require Different Architectures
- Speed of Innovation Outpacing Silicon Design Cycles

© Copyright 2018 Xilinx
Era of Pervasive Connected Intelligence

Adaptable Platforms Enable Rapid Deployment of Domain Specific Architectures (DSA’s) for Performance and Energy Efficiency

© Copyright 2018 Xilinx
The Imperative to Accelerate Innovation Deployment to Market

AI Papers Published

Source: Scopus

© Copyright 2018 Xilinx
The Era of Pervasive Intelligence Needs Adaptable Computing and DSA’s

Everything Intelligent & Connected

Deployed at Massive Scale

Dynamic Markets Need Rapid Innovation & Deployment
Whole Application Acceleration on Adaptable Platforms

Execution Time

- CPU Only
- CPU + ML ASIC
- CPU + FPGA
- App 1
- App 2
- App n

© Copyright 2018 Xilinx
Example: Smart / Safe City

**CPU-GPU**
- H.264 Decode
- Motion Analysis
- PCIe

**CPU**
- CNN

**GPU**
- Decode: 16ms
- OpenCV: 16ms
- CNN: 50ms
- > Power: 75W
- > Latency: 82 ms
- > Throughput: 4x12 fps

**CPU-Xilinx FPGA**
- H.264 Decode
- Motion Analysis
- PCIe

**CPU**
- CNN

**FPGA**
- Decode: 16ms
- OpenCV: 0.9ms
- CNN: 9.2ms
- > Power: 5W
- > Latency: 26.1 ms
- > Throughput: 4x38 fps

© Copyright 2018 Xilinx
Low-Latency, Highest Throughput Inference

- 30K images / sec on GoogLeNet
- AMD EPYC CPU
- Alveo U250 Accelerators
- Performance Scaling Up to 8 Accelerators

*batch=1, INT8

© Copyright 2018 Xilinx
RFSoC Leading the Way to 5G

- 16nm Zynq MPSoC Architecture with Integrated 4GSPS ADC, 6 GSPS DAC, and Programmable SDFEC

- Enables mMIMO with Dynamic Beamforming

- 7nm Versal with 10x Performance Increase for AI Beamforming and Cell Site Management
ACAP – The Next Big Advance in Adaptive Computing

Adaptive

Compute Acceleration

Platform

Scalar Engines

Adaptable Engines

Intelligent Engines

© Copyright 2018 Xilinx
Versal Prime Series
Broad Applicability Across Multiple Markets

- Mid-Range Series in the Versal Portfolio
- Optimized for Connectivity
- For In-Line Acceleration and Diverse Workloads

Versal AI Core Series
Breakthrough AI Inference Throughput

- Portfolio’s Highest Throughput for Low-Latency Inference
- Optimized for Cloud, Networking, and Autonomous Applications
- For Highest Dynamic Range of AI and Workload Acceleration
Platform for Any Developer

Adaptable for Any Application

Software Programmable

Heterogeneous Platform

User Application
C, C++, Python

Application-Specific Frameworks
- Machine Learning
- Video
- Genomics
- Search
- Financial Modeling
- Database

New Unified Software Development Environment

OS & Embedded Run-Time
- Custom HW
- Xilinx & Ecosystem HW Libraries
- C, Xilinx Libraries

Scalar Engines
- Adaptable Engines
- Intelligent Engines

VERSAL

© Copyright 2018 Xilinx
Low-Latency CNN Inference Performance

Sources:
- Alveo - Published (INT8)
- Versal - Projected (INT8), 65% PL reserved for whole application
- GPU 1 - P4 Published (INT8)
- GPU 2 - V100 Published (FP16/FP32)
- GPU 3 - T4 Projected

© Copyright 2018 Xilinx

Versal AI Core
Accelerating AI and Innovation
End-to-End in China

Xilinx @ China International Import Expo (CIIE)

Xilinx Developer Forum (XDF)

CIIE Shanghai, Nov. 5-10, 2018

XDF Beijing, Oct. 16, 2018

© Copyright 2018 Xilinx
Building the Adaptable, Intelligent World